Week 20 : Chi Square Distribution

pending tasks	
type	

Chi Square Distribution

- S^2 is used for sample variance. The expected value $E[S^2]$ of the sample variance and the variance $Var[S^2]$ of the sample variance can be found for given population parameters $\mu \: and \: \sigma$.
- Sample proportions $\hat{p}$ for a can be found from population proportion p for given Boolean conditions.
- In the previously seen example using Tesla cars, the mean mileage of the sample was used as the random variable, here the sample variance will be used as the random variable mapping each sample to the real number line.

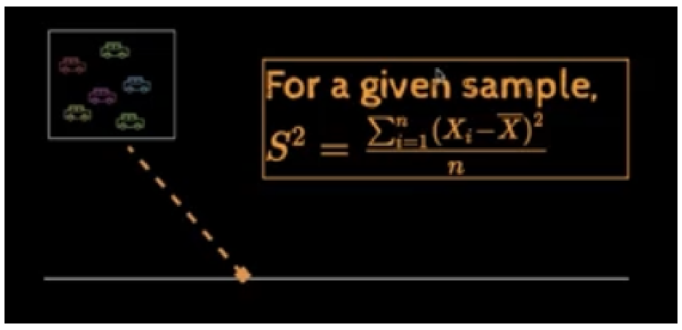


Fig.1 sample variance calculation. Note that the mean used is sample mean and not the population mean.

Estimating $E[S^2]$

Exercise(discrete) : Consider each sample as three dice throws. Sample 100,000 such samples, compute S^2 . Compute average of the 100,000 $S^2 values.\backslash$

Observations :
A discrete distribution of S^2 is obtained with average $S^2$of 1.944.
The population parameters are $\mu = 3.5$ and $\sigma^2 = 2.916$
It can be seen that $E[S^2] < \sigma^2$

Exercise(continuous) : Take a continuous distribution. Sample 3 values from $N(0,1)$. Compute S^2. Repeat for 100,000 samples.

Observations : The sample variance S^2 = 0.665 . This is also less than than the population variance σ^2.

Thus, $S^2 < \sigma^2$

Estimating $E[S^2]$ - Exercise

The variance is observed by tweaking the population parameters $\mu \: and \: n$ of the standard normal distribution.

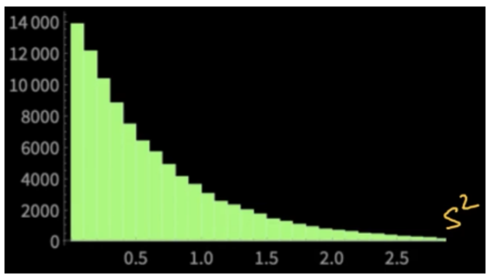


Fig.2a Distribution for altered population mean.

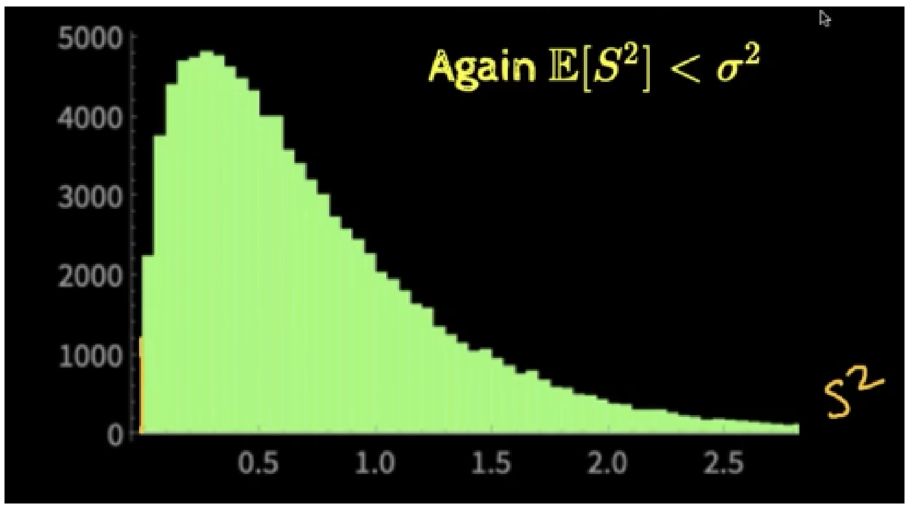


Fig.2b Distribution for altered sample size n.

1. For n = 3 and $\mu = 4$, a distribution similar to the previous one is obtained (Fig.2a). Thus, the inequality is unaffected. Average of $S^2 = 0.666$

2. For n =4 and $\mu = 0$, a change in distribution is observed(Fig.2b). Average of $S^2 = 0.749$.
   The density of the curve has moved to the right, thus increasing the spread with increase in n.

Exercise : Sample n values from N(0,1).

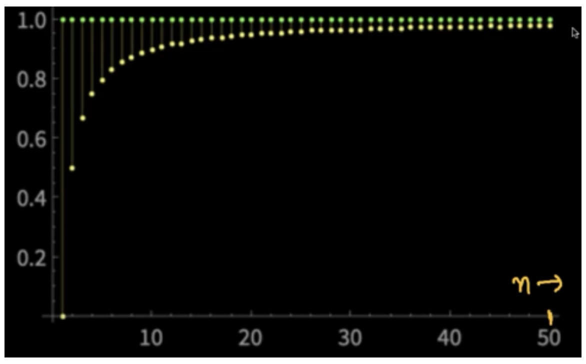


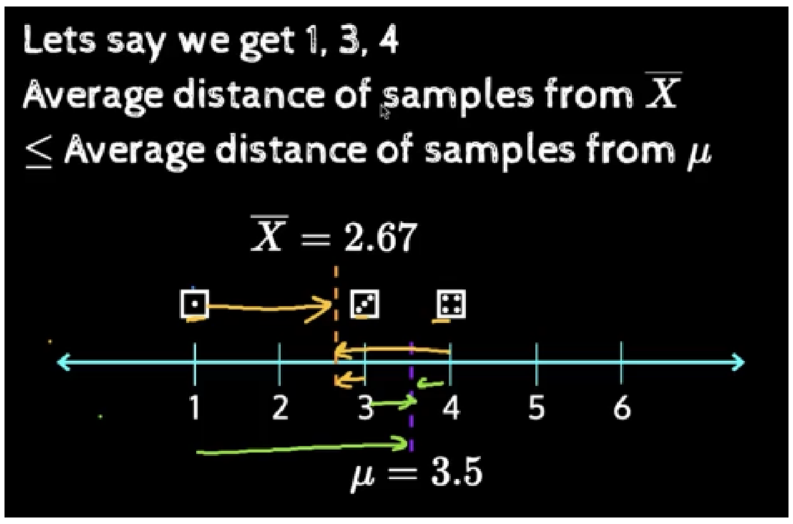
Fig. 3 plot of spread vs n.

Observation : The yellow dots, $E[S^2]$ move close to the green dots (the population variance $\sigma^2 = 1$)with the increase in n .

1. It can be hypothesized that $E[S^2] \to \sigma^2 as\ n \to \infty$.
2. $E[S^2]$ is always an **under-estimate** of σ^2 and gets more accurate as the sample size increases.
   The only difference is the calculation of $E[S^2]\ and\ \sigma^2$ is the mean value used. The under estimation occurs though the sample mean $\bar{X}$ is an unbiased estimate of population mean μ. i.e $E[\bar{X}] = \mu$.

Geometric Argument

- Population mean is always constant, for the example of dice it is 3.5. But the sample mean keeps varying. It is closer to the points if not equal to the population mean. Thus, the expected sample variance is always less then or equal to the expected population variance.



Lets say we get 1, 3, 4
Average distance of samples from $\bar{X}$
$\leq$ Average distance of samples from μ

Fig.4a The sample mean is closer to the sample points combined (when it is not equal to the population mean).

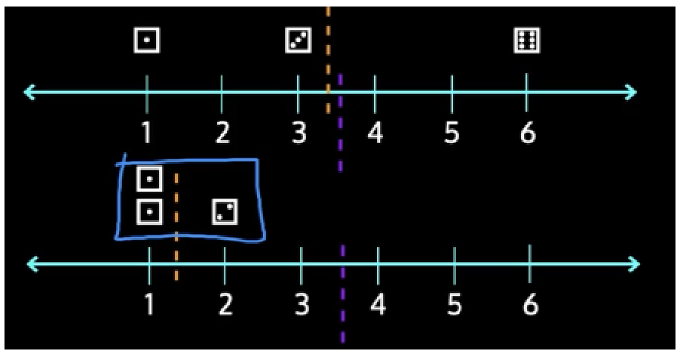


Fig.4b Examples of spread out and clustered samples.

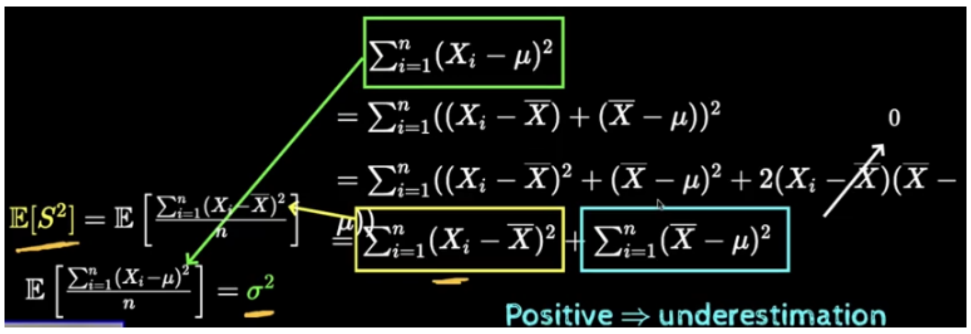
$\bar{X}$ will be close to μ in a well spread out sample.

For clustered samples, the sample mean can be significantly less than the population mean.

- Thus, in every case there is an under estimation of the sample variance.

Algebraic Argument

- Since the difference is only in the numerator of the variance calculation for sample and population, we derive answer in terms of the numerator.



$$\boxed{\sum_{i=1}^{n}(X_i-\mu)^2}$$

$$=\sum_{i=1}^{n}((X_i-\overline{X})+(\overline{X}-\mu))^2$$

$$=\sum_{i=1}^{n}((X_i-\overline{X})^2+(\overline{X}-\mu)^2+2(X_i-\overline{X})(\overline{X}-$$

$$\mu))$$

$$=\boxed{\sum_{i=1}^{n}(X_i-\overline{X})^2}+\boxed{\sum_{i=1}^{n}(\overline{X}-\mu)^2}$$

$$\mathbb{E}[S^2]=\mathbb{E}\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n}\right]$$

$$\mathbb{E}\left[\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}\right]=\sigma^2$$

Fig.5 Proof of underestimation.

- $\sum(\bar{X}-\mu)$ is equal to zero but $\sum(\bar{X}-\mu)^2$ is not zero . Thus, the sample mean is equal to the population mean but the sample variance is an underestimation of the population variance.

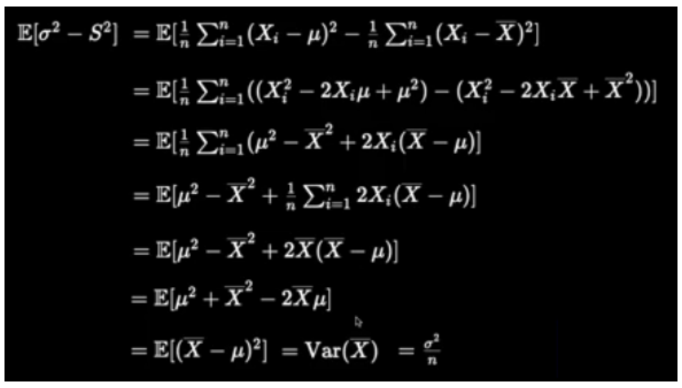
Find Expected Value of the Error

- Finding the expected value of error in variance in terms of μ, σ, n

It is known that

$$E[\bar{X}]=\mu$$
$$Var[\bar{X}]=\frac{\sigma^2}{n}$$
$$or$$
$$E[(\bar{X}-\mu)^2]=\frac{\sigma^2}{n}$$

The expected value of the error in variance $E[\sigma^2-S^2]$ can be found to be equal to $\frac{\sigma^2}{n}$.
It is done as follows (relating the sample variance to the sample mean , which is known):



$$\mathbb{E}[\sigma^2-S^2] = \mathbb{E}[\tfrac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^2-\tfrac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2]$$

$$=\mathbb{E}[\tfrac{1}{n}\sum_{i=1}^{n}((X_i^2-2X_i\mu+\mu^2)-(X_i^2-2X_i\overline{X}+\overline{X}^2))]$$

$$=\mathbb{E}[\tfrac{1}{n}\sum_{i=1}^{n}(\mu^2-\overline{X}^2+2X_i(\overline{X}-\mu)]$$

$$=\mathbb{E}[\mu^2-\overline{X}^2+\tfrac{1}{n}\sum_{i=1}^{n}2X_i(\overline{X}-\mu)]$$

$$=\mathbb{E}[\mu^2-\overline{X}^2+2\overline{X}(\overline{X}-\mu)]$$

$$=\mathbb{E}[\mu^2+\overline{X}^2-2\overline{X}\mu]$$

$$=\mathbb{E}[(\overline{X}-\mu)^2] = \mathrm{Var}(\overline{X}) = \tfrac{\sigma^2}{n}$$

Fig.6 Finding the expected error.

- It can be followed that

$$E[\sigma^2]-E[S^2]=\frac{\sigma^2}{n}$$
$$E[S^2]=\sigma^2(1-\frac{1}{n})=\sigma^2(\frac{n-1}{n})$$

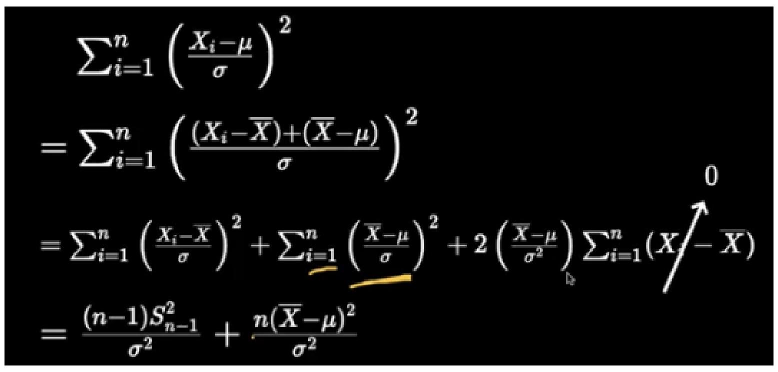
- The term $\frac{n-1}{n}<1$ and is the underestimation , which decreases as n increases.
- The sample variance is always smaller than the population variance. This is a systematic error and is called a bias. $E[S^2]$ is called a biased variance.
- An unbiased variance $S^2_{n-1}$is defined and $E[S^2_{n-1}]=\sigma^2$.
- From $E[S^2_n]=\frac{n-1}{n}\sigma^2$ and $E[S^2_{n-1}]=\sigma^2$ we can say

$$S^2_{n-1}=\frac{n}{n-1}S^2_n$$
$$S^2_n=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$$
$$S^2_{n-1}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}*\frac{n}{n-1}$$
$$S^2_{n-1}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$$

- The sample mean $\bar{X}$ and the sample variance (with a small correction) S_{n-1}^2 are unbiased estimates of the respective population parameters. Both hold, independent of the distribution function.
- To avoid confusion, for sample variance the unbiased variance value S_{n-1}^2 is used and for population variance S_n^2 is used. The difference can be significant for smaller values of n. And the formula used has to be checked while working with numberical packages.

Estimating $Var[S^2]$

- Sample variance is given by $S_{n-1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$, while using this for calculation of $Var(S_{n-1}^2)$, it would involve two RVs $X_i \ and \ \bar{X}$i.e. there two sampled quantities in each sum term.
- Because working with independent RVs is more convenient , the sample variance formula is rewritten to have only one samples value per term.
- $\sum_{i=1}^n (\frac{X_i - \mu}{\sigma})^2$ is an example of such a term where only X_i is the random variable.



$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma}\right)^2$$

$$= \sum_{i=1}^n \left(\frac{(X_i - \overline{X}) + (\overline{X} - \mu)}{\sigma}\right)^2$$

$$= \sum_{i=1}^n \left(\frac{X_i - \overline{X}}{\sigma}\right)^2 + \sum_{i=1}^n \left(\frac{\overline{X} - \mu}{\sigma}\right)^2 + 2\left(\frac{\overline{X} - \mu}{\sigma^2}\right) \sum_{i=1}^n \cancelto{0}{(X_i - \overline{X})}$$

$$= \frac{(n-1)S_{n-1}^2}{\sigma^2} + \frac{n(\overline{X} - \mu)^2}{\sigma^2}$$

Each term above contains only a single RV, although the above derivation is not useful for any generic population distribution.

- A Normal distribution is thus assumed for the RV X , $X \sim N(\mu, \sigma)$, since most of the real life data has normal distribution this is a plausible assumption.
- For a normal distribution, $(\frac{X_i - \mu}{\sigma})$ is the z-score of the term i. and the resultant terms have a standard normal distribution N(0,1).
- Thus, in $\sum_{i=1}^n (\frac{X_i - \mu}{\sigma})^2 = \frac{(n-1)S_{n-1}^2}{\sigma^2} + \frac{n(\bar{X} - \mu)^2}{\sigma^2}$, the LHS (Z_i) is the sum of many z-score values and the last term $(\bar{Z})$ in the RHS is a single z-score value, both having the standard normal distribution.

To understand the distribution of S_{n-1}^2 we have to find the distribution of sum of squares of standard normal variables(Chi Square Distribution).

Distribution of Sum of Squares of Standard Normal Variables

- Under the constraint that the population values are distributes normally,$X \sim (\mu, \sigma)$, the Chi square distribution is studied.
- Chi Square distribution is used to study standard normal variables. Let $Z_1, Z_2, ...Z_n$ be independent standard normal variables and $Q = \sum_{i-1}^n Z_i^2$.
- The following insights are derived from observing the Chi square distribution for n=1:

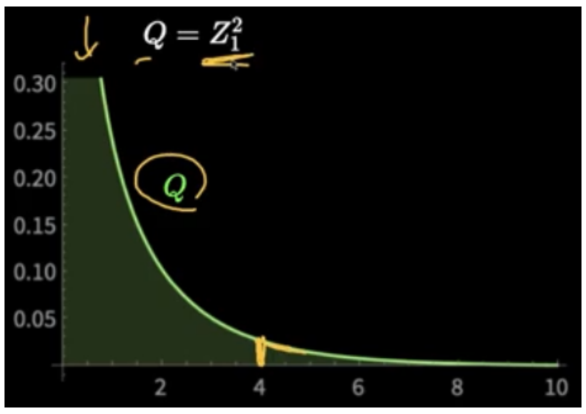


Fig.7 Chi square distribution for n=1.

1. PDF of Q for negative integers is 0 since it is a square term (Z^2).

2. Large numbers such as 100 can be obtained only when the Z value is close to 10, but in the standard normal distribution, occurrence of such numbers is low. Thus, farther away from the origin the value for PDF is low.
3. The PDF of given value, say 4 in Q will be greater than the standard normal distribution as it accounts for both 4 and -4 from the standard normal distribution. Thus, Q will have a longer tail than N(0,1).
4. For all values between (-1,1) the corresponding value for Q will be smaller, as $x^2 < x \, if \, x < 1$. This implies that the values will move closer to the origin, thus, the peak near origin will be higher for Q than N(0,1).

Distribution for N>1

def degrees of freedom - (t = 12:28)

- For n=2 , $Q = Z_1^2 + Z_2^2$, where both $Z_1 \, and \, Z_a$ are normally distributed. The negative half of Q is 0 and the tail is longer than for n=1, thus, there is more area further away from the origin.
- The area near the origin is lesser than the observed for n=1 because, probability of values of Q near the origin such as P(Q<0.01) requires that $P(q < 0.01) < P(|Z_1| < 0.1) * P(|Z_2| < 0.1)$ which has a fairly low chance of occurrence. And as n increases this further decreases, moving the peak away from the origin as seen for n=3 in fig.8b.

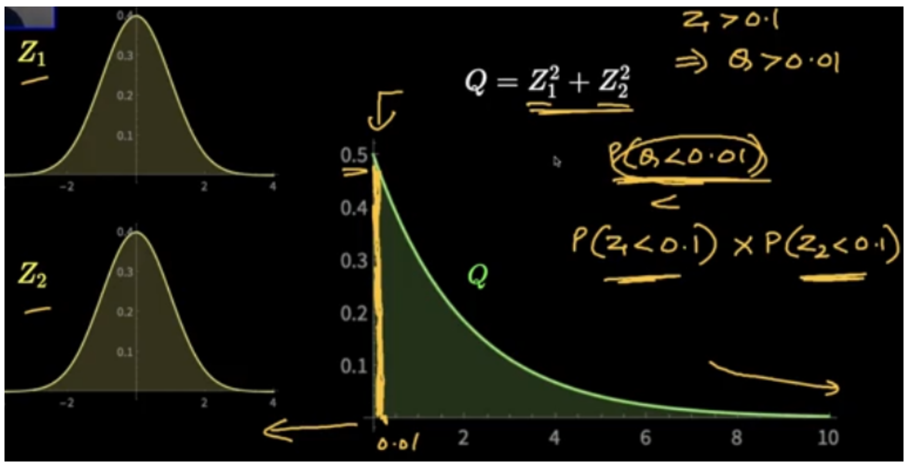


Fig.8a Distribution of Q for n=2

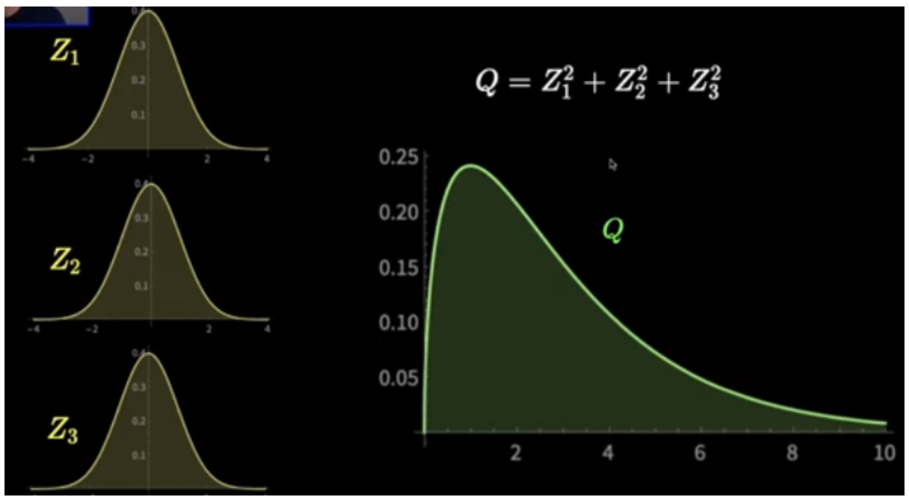


Fig.8b Distribution of Q for n=3.

- For n-1 and n=2 Q had monotonic distribution but n=3 is a non monotonic function and has a mode.
- Increasing n will move the mass towards the right, decrease the density towards the left and increase the spread.
- For large value of n, Q will be sum of multiple independent random variables and from CLT it follows that the resultant distribution will have a normal distribution. Here it is observed that the mean of the chi square distribution is equal to n.

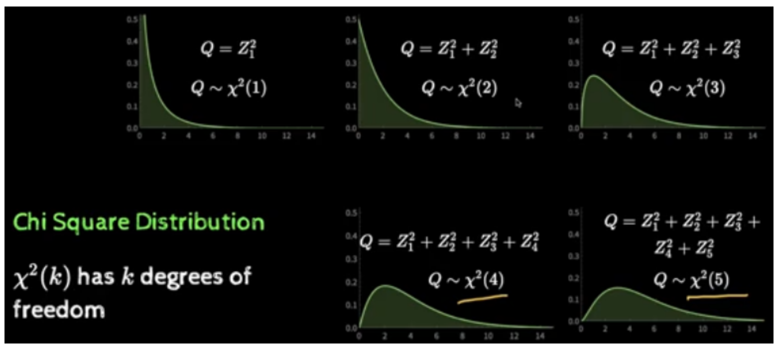


Fig.8c Examples of chi square distribution.

- These distributions, the sum of of square of k random variables is denoted by $\chi^2(k)$ where the distribution is said to have k degrees of freedom.

k Degrees of Freedom

- The mean of $\chi^2(1)$ can be found by using the relation $\mu_{\chi^2(1)} = E[Z^2]$, where $Z \sim N(0,1)$.This values can be derived from the relation
  $\sigma^2(Z) = E[Z^2] - (E[Z])^2$
  For the standard normal distribution, the mean is 0 and variance is 1, hence, $E[Z^2] = 1 \implies E[Q] = 1$

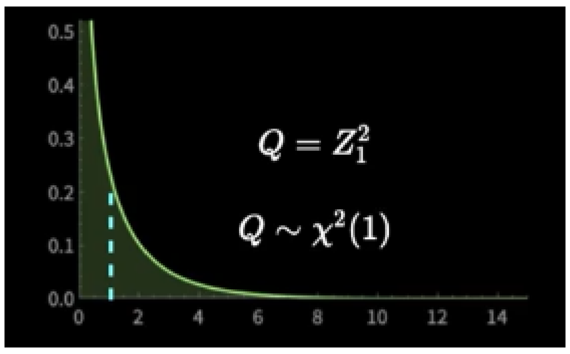


Fig.9a Chi square distribution for k=1.

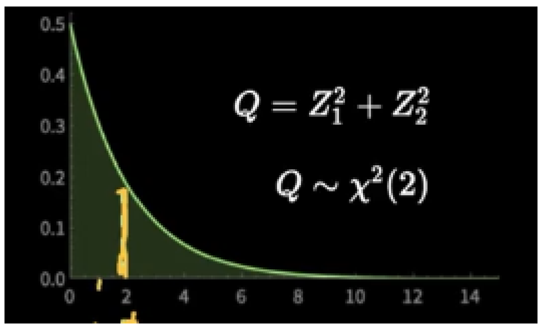


Fig.9b Chi square distribution for k=2.

- The mean of $\chi^2(2)$ can be found using $E[Q] = E[Z_1 + Z_2]\ or\ E[Z_1^2] + E[Z_2^2]$ since, $Z_1\ and\ Z_2$ are independent and is found to be 2. This implies that more area has shifted to the right.
- Thus, mean of $\chi^2(k) = k$.i.e mean of chi square is equal to it's degree of freedom. With increasing degree of freedom k , the density and the mean move to the right.

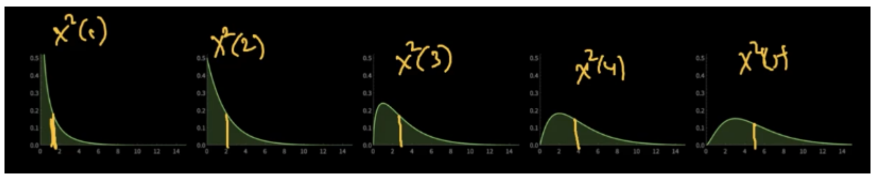


Fig.9c Observing the pattern between mean of chi square distribution and k.

Variance of $\chi^2(k)$

def chi square distribution - (t = 6:12)

- The variance of $\chi^2(1) = \sigma^2(Z^2)$
  $\sigma^2(Z^2) = E[Z^4] - (E[Z^2])^2$
  The second term $(E[Z^2])^2 = \mu(\chi^2(1)) = 1$. For the first term , it can be followed from CLT (moments of distribution) that the n^{th} moment is given by the expected values of the random variable raised to n. In this case, it is the fourth moment of the standard normal distribution.

  The fourth moment is called **Kurtosis.**

> **Recap:** that the first moment was mean, second was the variance and the third moment relates to skewness.

It can be computed with the integral $\int_{-\infty}^{+\infty} x^4 f(x)dx$, where $f(x)$ is the PDF of $N(0,1)$. The area under this curve is 3.

Thus, $\sigma^2(Z^2) = 3 - 1 = 2$, is the variance of Z^2.

.i.e $var(\chi^2(1)) = 2$

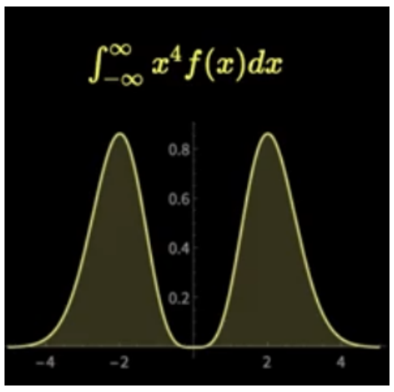


Fig.10 Curve used to calculate the fourth moment.

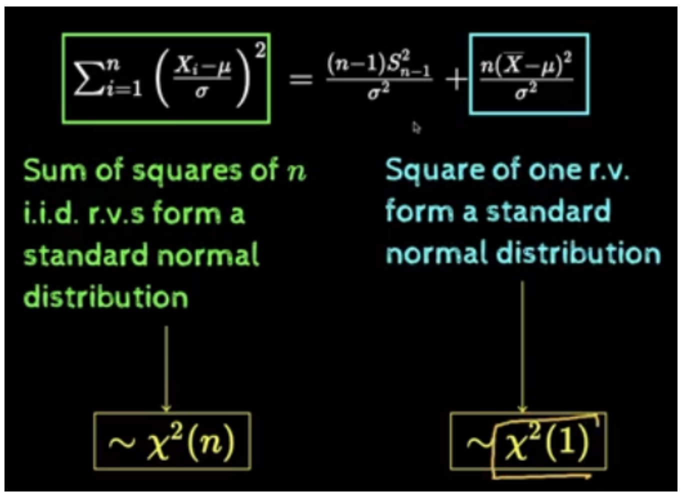
- Similarly, for $Q = Z_1^2 + Z_2^2$, (Z_1, Z_2 are independent)

$$var(Q) = var(Z_1^2 + Z_2^2)$$
$$= var(Z_1^2) + var(Z_2^2) = 2 + 2 = 4$$

- It can be extrapolated that, $var(\chi^2(k)) = 2k$ for a given k. So the mean of chi square is k(the distribution is moving rightwards with the addition of new variables) , while the variance is 2k (increasing variance makes the curve more flatter with increasing k).
- **$\chi^2(k)$ is the distribution of the sum of the squares of k standard normal distributions. k is referred to as the degrees of freedom. Mean of $\chi^2(k)$is k and the variance is 2k.**

- Chi square distribution is useful in the estimation of confidence intervals for a given variance. It is also used to quantify the goodness of fit and testing the independence of two variables.

Recap & Statistics of S^2



$$\boxed{\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2} = \frac{(n-1)S_{n-1}^2}{\sigma^2} + \boxed{\frac{n(\bar{X}-\mu)^2}{\sigma^2}}$$

Sum of squares of n i.i.d. r.v.s form a standard normal distribution

$\boxed{\sim \chi^2(n)}$

Square of one r.v. form a standard normal distribution

$\boxed{\sim \chi^2(1)}$

Fig.11 Calculating sample variance.

- The term in the middle (in the above figure) is the unbiased sample variance. It is surrounded by by two terms with a chi square distribution on the LHS and RHS with n and 1 degree(s) of freedom respectively.

- By equating the LHS, RHS and using some uniqueness theorems it can be argued that the middle term has (n-1) degrees of freedom i.e. has a distribution $\sim \chi^2(n-1)$. $\frac{n-1}{\sigma^2}$ is a constant therefore, S_{n-1}^2 has a χ^2 distribution with n-1 degrees of freedom.

- S_{n-1}^2 has n terms adding up, but they are not independent. Since, in $\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ the linear term $\sum_{i=1}^{n}(X_i-\bar{X})=0$. This implies that if n-1 terms are freely chosen then the last one gets fixed. Thus, one degree of freedom is lost.

- The mean of sample variance $E[S_{n-1}^2]=\sigma^2$ is an unbiased estimate of the population variance. This relation can be checked as follows:

$$\begin{aligned} E[\frac{(n-1)S_{n-1}^2}{\sigma^2}] &= Mean\ of\ \chi^2(n-1) \\ \frac{(n-1)}{\sigma^2}E[S_{n-1}^2] &= n-1 \\ E[S_{n-1}^2] &= \sigma^2 \end{aligned}$$

- The variance is calculated as follows

$$\begin{aligned} Var[\frac{(n-1)S_{n-1}^2}{\sigma^2}] &= Var\ of\ \chi^2(n-1) \\ (\frac{(n-1)}{\sigma^2})^2 Var[S_{n-1}^2] &= 2(n-1) \\ Var[S_{n-1}^2] &= \frac{2\sigma^4}{n-1} \end{aligned}$$

This implies, the greater the variance of the population the larger would be the sample variance (grows by σ^4).

- In conclusion, the sample statistic S_{n-1}^2 defined as $\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$ is an unbiased estimator of the population variance σ^2, i.e. $E(S_{n-1}^2)=\sigma^2$. This result for expected value of variance holds for all distributions.

- But the expected value of variance holds true only if the population values are normally distributed. i.e. if the population values are normally distributes, then $\frac{(n-1)S_{n-1}^2}{\sigma^2} \sim \chi^2(n-1)$ and $Var[S_{n-1}^2]=\frac{2\sigma^4}{n-1}$.

On to Experiments

- The sample size of n corresponds to chi square distribution with freedom n-1.

- If the assumption that the underlying distribution is normal does not hold true then the calculation of sample distribution and variance using chi square does not hold true, especially with small sample size. As the sample size is increased, a normal distribution is obtained.

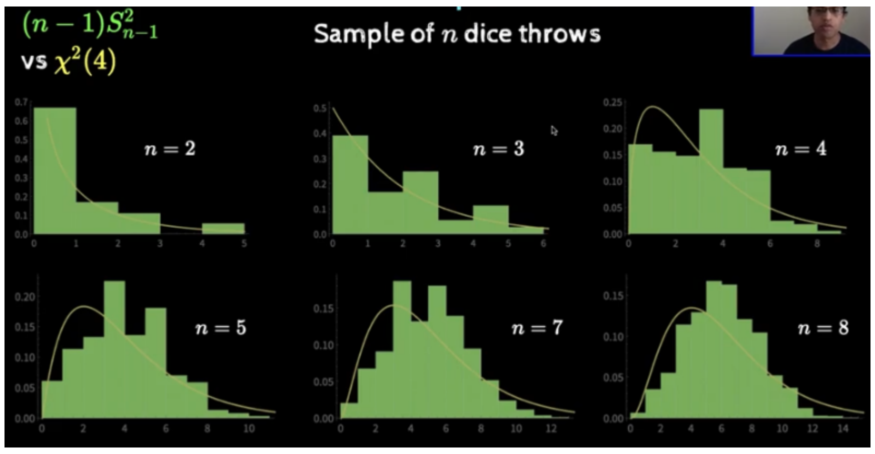


Fig.12 sample distribution vs chi square distribution for n dice throws.

Expectation of Proportion

def indicator variable - (t = 6:12)

- Calculating expected value and variance of the sample proportion.
- Taking the previously discussed example of finding king of diamonds in a set of 13 cards from a deck of 52 (where the boolean proportion is found), we find how sample proportion is related to the population proportion.
- Let p be the proportion of elements in the population satisfying a condition and $\hat{p}$ be the proportion of elements in a sample satisfying a condition.
- An **Indictor variable** (a type of random variable) indicates weather a given boolean condition is true or false. More formally, define a random variable X such that $X_i = 1$, if ith element of sample satisfies condition, 0 other wise.
- $\hat{p}$ is sample mean of the indicator random variable X.

$$\hat{p} = \frac{X_1 + X_2 + ... + X_n}{n} = \bar{X}$$

Also, taking the same indicator RV for the population gives

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N} = p$$

i.e. the population parameter mean is equal to population parameter proportion. it is already known that $E[\bar{X}] = \mu$. hence,

$$E[\hat{p}] = E[\bar{X}] = \mu = p$$

Sample statistic of proportion is an unbiased estimate of population parameter of proportion.

Variance of Proportion

Variance of $\hat{p}$ is equivalent to variance of the sample mean $\bar{X}$.

The population parameter σ^2 for the indicator variable is given by

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X-\mu)^2}{N}$$
$$= \frac{pN(1-p)^2 + (1-p)N(0-p)^2}{N}$$
$$= p(1-p)$$

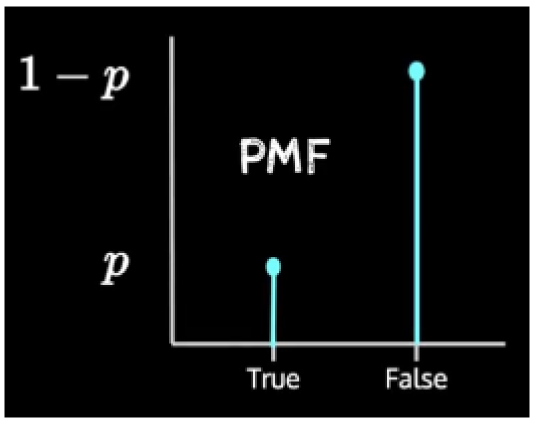


Fig.13 PMF of RV X.

Thus, standard deviation $\sigma = \sqrt{p(1-p)}$.

$$var(\bar{X}) = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = var(\hat{p})$$

- The variance is largest when p =0.5, $var(\hat{p}) = \frac{0.25}{n}$ and would require larger sample sizes. The variance falls linearly as the number of elements in the sample is increased.